Análise Exploratória e Experimentação de Diferentes Métodos de Tratamento de Dados Omissos

Bingnan Zheng
Coimbra Business School | ISCAC, Politécnico de Coimbra
Coimbra, Portugal
a2019130038@alumni.iscac.pt

Carolina Freire
Coimbra Business School | ISCAC, Politécnico de Coimbra
Coimbra, Portugal
2021103980@alumni.isca.pt

Raquel Miranda
Coimbra Business School | ISCAC, Politécnico de Coimbra
Coimbra, Portugal
a2022109542@alumni.iscac.pt

Resumo — **Dados em falta é um dos principais fatores que afetam a qualidade e o conhecimento eventualmente extraído de conjuntos de dados. No entanto, existem técnicas de imputação de dados ausentes (MDITs) que podem ser utilizadas para melhorar a qualidade dos dados do conjunto. Neste trabalho, apresentamos os conceitos referentes a dados em falta, os métodos usados para contornar esse problema de modo pratico, conforme se verá abaixo.**

Palavras Chave – Dados em falta; Análise exploratória; Padrões de dados omissos; MCAR; MAR; MNAR.

I. INTRODUÇÃO

Com o aumento significativo de dados coletados atualmente, a qualidade da descoberta de conhecimento extraido desses dados por meio de técnicas específicas depende da integralidade do conjunto de dados. Os dados omissos são os dados que estão em falta em determinado agrupamento, se trata de um problema generalizado, visto que eles escondem informações que podem vir a ser importantes na descoberta de conhecimento. Aliás, as omissões habitualmente ocorrem de modo padronizado, que são categorizadas como: omissos completamente ao acaso (MCAR), omissos aleatoriamente (MAR) e os omissos não aleatoriamente (MNAR). Desta forma, os dados omissos podem causar diversas intercorrências em uma análise de dados, como a perda da eficiência e fidedignidade estatística, redução significativa no tamanho da amostra, além de tornar os métodos comuns de análise de dados inapropriados ou difíceis de aplicação Para contornar esses problemas, foram desenvolvidas algumas técnicas de tratamento de dados omissos, sendo que os mais comuns são a taxonomia, a remoção e a imputação.

II. PADRÕES DE DADOS OMISSOS

A presença de dados omissos é comum em muitos conjuntos de dados, seja porque algumas informações não foram coletadas, porque os dados foram perdidos ou porque ocorreram erros de entrada de dados. Geralmente, os dados omissos estão ordenados em dois tipos de padrões, que auxiliam na escolha do método de dados omissos utilizado para tratar deste problema, visto que alguns métodos de dados omissos são específicos aos padrões de dados omissos.

A. Univariado vs. Multivariado

Os padrões de dados omissos podem ser classificados como univariados ou multivariados, dependendo de como os dados omissos estão distribuídos nos diferentes atributos ou variáveis do conjunto de dados.

No padrão univariado, apenas uma variável do conjunto de dados está em falta. Por exemplo, pode haver um conjunto de dados em que as informações sobre a idade de diversos respondentes não estão disponíveis, mas todas as outras informações pessoais, como sexo, nível de educação e renda, estão completas.

Por outro lado, no padrão multivariado há dados em falta em mais de uma variável do conjunto de dados. Por exemplo, pode haver um conjunto de dados em que informações sobre a idade e o nível de educação dos respondentes estejam em falta. Nesse caso, o padrão de dados omissos é multivariado, porque há dados faltantes em duas variáveis diferentes. Este padrão é geralmente mais complexo e desafiador de lidar do que o padrão univariado, uma vez que a ocorrência de dados omissos em várias variáveis pode afetar as relações e correlações entre elas. Nessas situações, a abordagem de preenchimento de valores ausentes com a média ou mediana pode introduzir viés ou distorcer a distribuição dos dados, e técnicas mais sofisticadas de modelagem, como a imputação múltipla, podem ser necessárias para lidar com os dados omissos de forma mais precisa e robusta.

B. Relevância da Visualização

A visualização dos padrões de dados ausentes é fundamental na análise de dados, uma vez que pode ajudar a identificar possíveis padrões e tendências nos dados ausentes e fornecer informações valiosas para lidar com esses dados de forma apropriada. Pode revelar se os dados faltantes são aleatórios ou se há algum padrão entre os dados ausentes.

Esta visibilidade dos dados pode ajudar a identificar variáveis que têm uma proporção significativa de dados faltantes, o que pode ser útil para decidir se essas variáveis devem ser incluídas ou excluídas da análise, além de também auxiliar na identificação de possíveis erros ou problemas nos dados de entrada que levaram à omissão de dados.

Existem várias ferramentas de visualização disponíveis para representar padrões de dados omissos, como gráficos de barras, gráficos de dispersão e mapas de calor. Essas visualizações podem ser úteis para entender a distribuição e a proporção dos dados ausentes em cada variável, e para comparar a distribuição dos dados ausentes entre diferentes subgrupos de dados.

C. *Referência à percentagem de dados em falta*

Alguns métodos de imputação de dados podem não funcionar corretamente com uma percentagem alta de dados omissos. A percentagem aceita de modo geral, por muitos, para que um conjunto de dados seja considerado utilizável é de até 40% de dados em falta. Entretanto, isso depende do contexto e do objetivo da análise de dados. Em alguns casos, mesmo uma pequena percentagem de dados faltantes pode ser problemática, enquanto em outros, uma percentagem maior pode ser tolerável.

III. Mecanismos/Tipos de Dados Omissos e sua Deteção

Tendo os conceitos acima esclarecidos em mente, passamos a efetiva solução do problema de dados incompletos em determinado conjunto de dados. O primeiro passo para a solução dessa dificuladade é a identifição do mecanismo que corretamente descreve a distribuição dos valores omissos e sua implicação na inferência estatística.

A. *Mecanismos de Dados Omissos*

Mecanismos de dados omissos referem à forma como os dados em falta estão distruibuidos nas variáveis do estudo. O reconhecimento do mecanismo baseia-se na aleatoridade da distribuição dos valores omissos nas variáveis, podendo ser classificados como: Omissos completamente ao acaso (MCAR); Omissos Aleatoriamente (MAR); e Omissão não aleatoriamente (MNAR):

1) Dados Omissos Completamente ao Acaso (Missing Completely at Random - MCAR): Trata-se de casos em que os dados em falta não estão relacionados com qualquer das variáveis, constituindo uma amostra verdadeiramente aleatória. Neste caso, os dados que faltam não possuem nenhuma relação com a probabilidade de resposta e não existe viés ou tendência associados com o seu carácter aleatório.

2) Dados Omissos Aleatoriamente (Missing at Random - MAR): Aqui os dados faltantes não respeitam uma distribuição completamente aleatória, como na descrição anterior, havendo alguma relação entre os dados omissos e uma informação contina em outra variáveis que não contenha dados em falta. Ou seja, os dados omissos estão relacionados a outras variáveis do estudo, mas não estão relacionados com os próprios dados. Neste caso, a probabilidade de resposta está relacionada com outras variáveis que já foram observadas ou medidas, de tal forma que é possível construir modelos para prever os dados que faltam.

3) Dados Omissos Não Aleatoriamente (Missing Not at Random - MNAR): Na terceira e última classificação, o padrão dos dados omissos não é aleatório, sendo provável que os valores faltantes dependam não só dos dados *observados, mas* também dos dados não observados. Uma importante dificuldade dessa classificação é que não é possível identificá-lo sem saber previamente o valor que está ausente.

De tudo, têm-se que o conhecimento do mecanismo de dados omissos é fundamental para selecionar a técnica de imputação mais apropriada e avaliar os efeitos potenciais dos dados omissos na análise de dados.

4) Deteção de Dados Omissos: A perfomance dos métodos de dados omissos depende do mecanismo subjacente de dados ausentes. Se a relação entre a probabilidade de dados ausentes e as variáveis observadas não puder ser detetada e não havendo motivos para os dados estarem em falta, assume-se que os dados são MCAR. Caso contrário, podem ser MAR ou MNAR. É de suma importância averiguar os motivos pelos quais os dados estão em falta.

A deteção de dados omissos envolve a identificação e o tratamento de valores ausentes em um conjunto de dados. Existem várias maneiras de detetar dados omissos, alguns métodos comuns incluem:

- Visualização de dados: Uma das maneiras mais simples de detetar dados omissos é através de visualização de dados, como gráficos de barras, histogramas ou gráficos de dispersão. Os dados omissos aparecem como lacunas nos gráficos.
- Análise de frequência: Pode-se usar uma análise de frequência para verificar a frequência de dados ausentes em cada variável.
- Análise de correlação: Pode ajudar a detetar padrões de associação entre variáveis que contenham valores ausentes.
- Algoritmos de deteção de valores ausentes: Existem algoritmos específicos projetados para detetar valores ausentes em um conjunto de dados. São eles: Regras de decisão, Análise de componentes principais (PCA), Árvores de decisão, e algoritmos de agrupamento (clusterização).

a) Testes Estatísticos: Ajudam a determinar se os dados faltantes em um conjunto de dados estão relacionados a outras variáveis ou se eles foram perdidos aleatoriamente, ou seja, se é o mecanismo é MCAR ou não-MCAR. Esses testes ajudam a reconhecer padrões ou razões para a omissão de dados. Existem diferentes tipos de testes estatísticos que podem ser usados para detectar a presença de dados faltantes relacionados. Algumas das técnicas mais comuns incluem:

- *Teste do chi-quadrado (Chi-square test):* Usado para verificar se a falta de dados está relacionada a uma ou mais variáveis do conjunto de dados. É útil quando há uma quantidade significativa de dados faltantes.
- *Teste t de comparação de médias(t-test procedure):* É usado para comparar as médias de grupos de dados que têm valores faltantes. Ele é útil quando os dados ausentes são distribuídos aleatoriamente ou estão relacionados a uma única variável.

- *Análise de correlação*: É usado para verificar se existe uma correlação entre os dados faltantes e outras variáveis do conjunto de dados.

IV. TRATAMENTO

Após a identificação do mecanismo e do padrão dos dados omissos, passamos ao tratamento desses dados para que o conjunto de dados possa ser utilizado em processos de descoberta de conhecimento.

Hoje em dia existem variados métodos para o tratamento dos dados em falta (*Missing Data Techniques - MDTs*), que geralmente são divididos em três categorias e que serão estudados a seguir: remoção, imputação e tolerância.

A. Remoção

A remoção de dados faltantes é uma abordagem relativamente simples, mas pode levar a uma perda significativa de informações. Consiste em unicamente excluir os casos que possuem dados em falta. Apesar de muito utilizado, não é o método mais eficiente e deve apesar ser usada quando a quantidade de dados omissos é muito pequena. Essa técnica possui duas formas:

1) Análise de Casos Completos (Complete Cases ou Listwise Deletion - LP): Esse método deleta os casos que contêm dados em falta, usando para a análise somente os casos que possui todos os dados presentes. Uma de suas desvantagem é que sua aplicação leva a uma grande perda de observações, além de induzir a conclusão obtida, visto que o método não lida corretamente com conjuntos de dados incompletos quando os valores não estão faltando completamente ao acaso (MCAR).

2) Análise de Casos Disponíveis (Available Cases Analysis ou Pairwise Deletion - PD): A técnica considera cada variável separadamente, sendo que todos os valores são considerados e os dados em falta são simplesmente ignorados, o que significa que a análise lida com uma amostra desigual para cada uma de suas variáveis, o que pode dificultar a interpretação do resultado. Esse método somente não induzirá um resultado quando os dados estiverem em MCAR.

B. Imputação

Existem várias técnicas de imputação de dados que podem ser usadas para tratar dados omissos, cada uma com suas vantagens e desvantagens. A estratégia consiste basicamente na inferencia dos dados omissos a partir dos dados presentes. Esses métodos podem ser classificados da seguinte maneira:

1) Classificação dos Métodos de Imputação:

a) Univariada vs. Multivariada: A técnica Univariada usa apenas os valores não omissos de determinada variável e atribui um valor aos dados em falta dessa variável utilizando uma constante escolhida pelo usuário ou utilizando uma medida estatística (média) da variável. Por outro lado, a técnica Multivariada usa todo o conjunto de dados (linhas e colunas) para prever o valor em falta em determinada variável. A última é uma solução mais sofisticada.

b) Hot-deck vs. Cold-deck: Hot-deck substitui os valores em falta pela observação mais semelhante com base nas outras variáveis do conjunto de dados. Uma de suas desvantagens é a definição do que é similar, o que descarta a importância de realizar a comparação entre diferentes técnicas de agrupamento. *Cold-deck* é similar à técnica *Hot-deck*, mas utiliza outro conjunto de dados para substituir os dados em falta do cojunto em análise; sendo útil para variáveis que são estáticas.

c) Única vs. Múltipla: A técnica Única se baseia na substituição do valor em falta por um valor único, sem a utilização de um modelo fixo, mas com dados prédeterminados. É benéfico por sua simplicidade e por lidar com todos os dados omissos apenas uma vez, tornando o conjunto de dados passível de análise sem qualquer dados em falta. Já a Múltipla, é ideal para um conjunto MAR, vez que se aproveita da criação de vários conjuntos de dados plausíveis de serem imputados e combina os resultados obtidos de cada um deles. A imputação múltipla difere dos métodos de imputação única porque os dados ausentes são preenchidos muitas vezes, com muitos valores plausíveis diferentes estimados para cada valor ausente.

*2) Imputação Por Uma Medida De Tendência Central Ou Outra Constante:*Trata-se da imputação de um dados ausente pela média dos dados não ausentes daquela variável, de conjuntos de dados MCAR. Este método mantém o tamanho da amostra e é fácil de usar, mas a variabilidade nos dados é reduzida, de modo que os desvios padrão e as estimativas de variância tendem a ser subestimados. Em contraponto, o método leva a um viés em conjuntos multivariados, como correlação ou coeficientes de regressão.

3) Imputação Por Regressão Linear: Este método usa um modelo de regressão linear para prever os valores faltantes com base nas outras variáveis do conjunto de dados. Inicialmente o modelo de regressão é estimado nos dados observados e, subsequentemente, usando os pesos da regressão, os valores ausentes são previstos e substituídos. Vale esclarecer que existem diversas configurações do modelo de regressão linear para obtenção do valor pretendido, de modo que todas possuem vantagens e desvantagens a depender do caso concreto. Essa técnica pode ser mais precisa do que as técnicas simples de imputação, mas pode ser computacionalmente intensiva e pode levar a problemas se houver multicolinearidade entre as variáveis.

4) Imputação Por K-Vizinhos Mais Próximos (K-Nearest Neighbours - k-NN): Esse método emprega um cálculo pairwise de uma certa distância ou medida de similaridade entre os valores presentes na variável para imputar os dados omissos. Essa técnica é semelhante à imputação por *hot-deck*, mas usa várias observações em vez de apenas uma.

Essas são apenas algumas das técnicas de imputação disponíveis, e a escolha da técnica depende do conjunto de dados e do objetivo da análise. É importante lembrar que a imputação de dados pode ter um impacto significativo nos resultados da análise, portanto, é essencial escolher cuidadosamente a técnica de imputação e considerar a

possibilidade de que os valores imputados possam introduzir viés na análise, conforme se visualiza na Fig. 1.

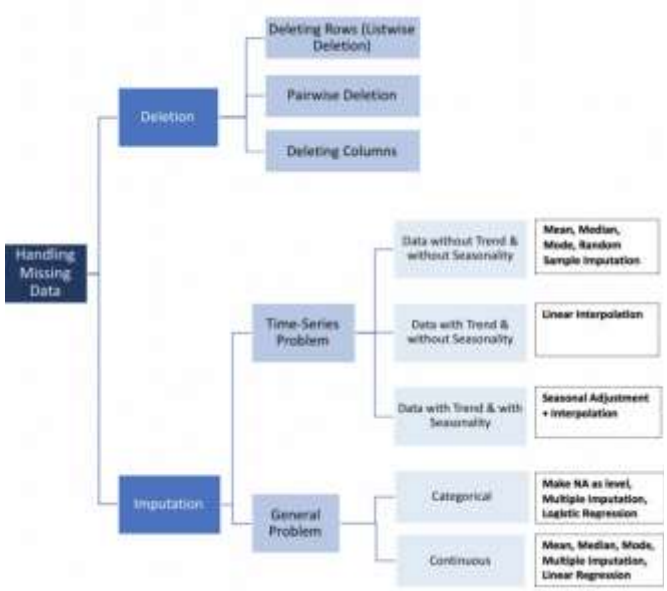

Figura 1. https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

C. *Tolerância*

As técnicas de tolerância são uma boa opção quando o objetivo não é prever os dados em falta. A estratégia desta técnica é baseada no tratamento interno, sendo que os dados em falta no conjunto de dados são tolerados e a análise é realizada diretamente no conjunto de dados. Um desses tipos de abordagem de tolerância é atribuir um valor nulo para substituir a parte ausente dos dados

V. Resultados E Discussão

No presente trabalho, foi solicitado ao grupo a realização de uma análise exploratória e a experimentação de diferentes técnicas de imputação de dados omissos no conjunto de dados de nº 2, designado pela Prof. Joana Leite. A análise exploratória foi efetuada admitindo que o propósito final do estudo seria o de prever o preço de venda dos imóveis descritos no conjunto de dados. A seguir, passamos a elucidar algumas das técnicas utilizadas no notebook elaborado em complementação ao presente artigo e entregue em anexo.

Antes de tudo, é necessário aferir se existem dados omissos no conjunto de dados. No conjunto estudado, cada linha simboliza um imóvel e cada coluna, as caracteristicas relevantes deste imóvel, como número de quartos, número de banheiros, ano de construção e outros. No conjunto, as colunas que apresentaram dados omissos foram são: grade (classificação de mercado do apartamento), sqft_basement (metros quadrados do porão) e zipcode (código postal). Além disso, é preciso aferir os dados com valor zero, para que certificar que não há dados omissos 'escondidos'. Desta forma, verificou-se que as colunas bedrooms e yr_renovated possuem dados com valor zero, o que é normal para o caso, não se tratando de dados omissos.

Como comentado anteriormente, a percentagem dos dados omissos é um fator importante na decisão do tratamento de dados e na sequente imputação de dados. Na situação em análise, das três colunas com dados omissos, o zipcode apresenta 4,44% de valores omissos, o *grade* 28,51% e o mais grave é o sqft_basement, que apresenta 63,34%:

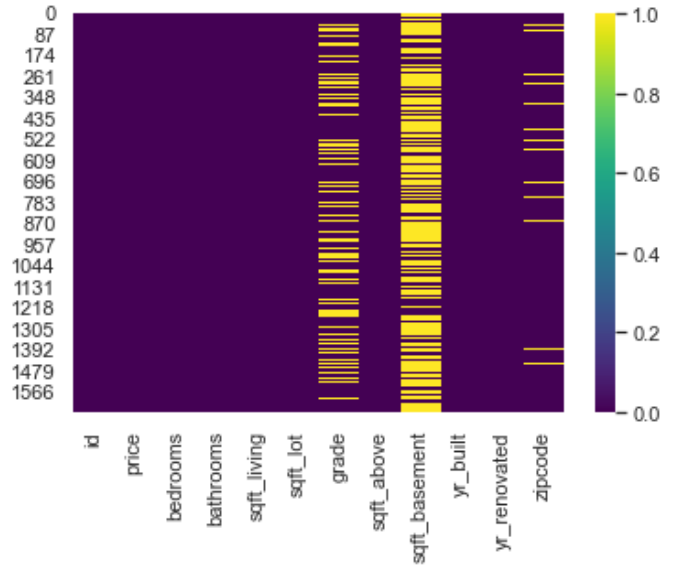

Figura 2. Autoria própria.

Com isso em mente, passamos a exploração individual de cada uma dessas colunas.

A. *Grade*

Nesta coluna, os dados em falta seguem uma distribuição normal, isso é, tem como valor mais alto um ponto de classificação mediana e em seguida decresce para 'os dois lados', como demonstra a Fig. 3. Isso significa que os valores em falta possuem correlação com outros valores da tabela e que podem ser classificados como Dados Omissos Aleatoriamente (MAR):

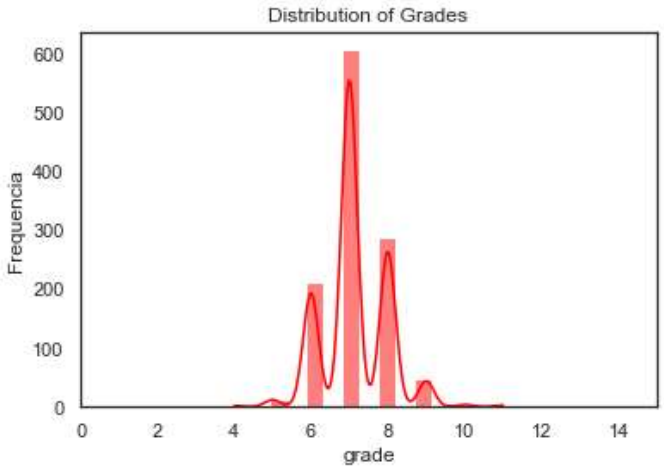

Figura 3. Autoria própria.

Considerando que o *grade* dos imóveis pode possuir alta correlação com os outros valores do conjunto de dados, foi realizado o estudo dessa correlação. Neste cálculo foi possível

evidenciar que em verdade o *grade* não possui alta correlação com outro indicador, pois a correlação apresentada mais elevada é de 0,59, com o *sqft_above:*

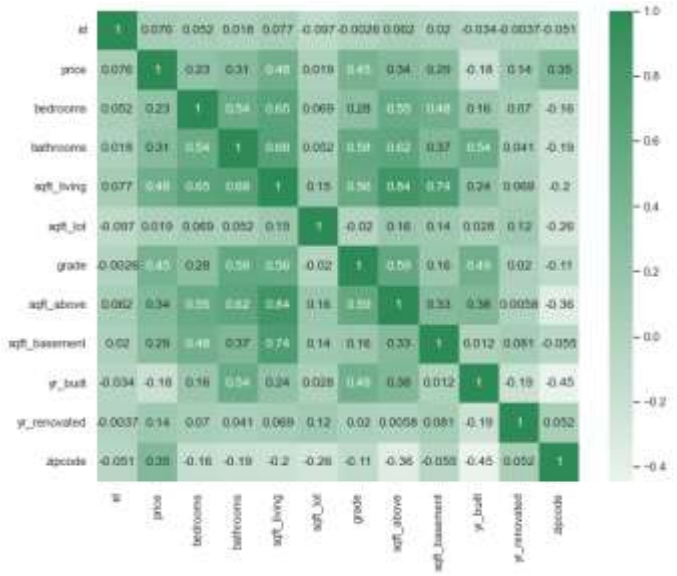

Figura 4. Autoria própria.

Como as correlações são em geral baixas, isso significa que um único indicador não é suficiente para determinar os dados omissos da coluna *grade*, atendendo às caraterísticas da coluna, buscamos uma solução usando a regressão linear. Assim, ante de preencher os as linhas com dados em falta, foi importante analisar a sua credibilidade. Como já havíamos determinado a distribuição normal dos dados não omissos, também é possível verificar a distribuição dos dados previstos pelo modelo.

Desta forma, foi possível aferir que os dados previsto pelo modelo também apresentam uma distribuição normal, divergindo dos dados não omissos apenas no sentido que os valores são levemente mais elevados:

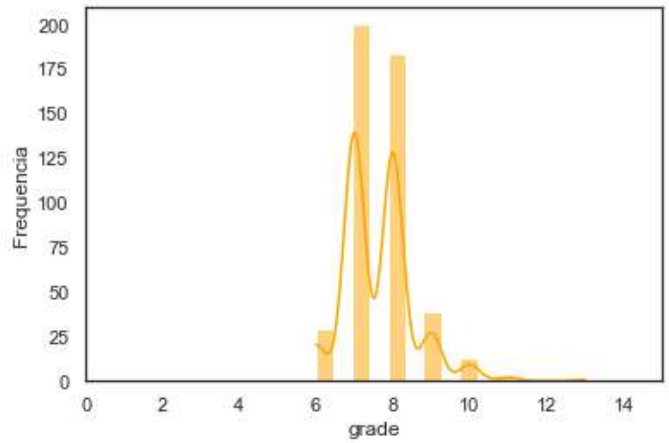

Figura 5. Autoria própria.

Dessa maneira, os dados previstos pelo modelo foram preenchidos na tabela e por fim, foi realizado um teste do R-quadrado para avaliar a credibilidade do modelo, de modo que as 5 variáveis que treinaram esse modelo conseguem explicar 61,34% da classificação da casa.

B. Sqft_basement

Não é particularmente difícil de reparar que todos os dados omissos desta coluna representam, em verdade, os imóveis que não possuem porão, ou seja, são omissões que representam zero. Desta forma, os dados em falta são Dados Omissos Não Aleatoriamente (MNAR).

A solução para estes dados é simples, basta atribuir o valor zero para todas linhas que estão sem dados. Mas essa solução pode trazer algum erro ao programa, sendo mais eficiente realizar essa transição por meio de um cálculo. Como existem as colunas dos valores da área do sqft_*living* (metros quadrados totais do imóvel) e do *sqft_above* (metros quadrados do imóvel sem o porão), através de uma subtração simples é possível alcançar o valor do *sqft_basement.*

C. Zipcode

A coluna *zipcode* difere das demais no sentido de, apesar de ser composta por valores numéricos, é um dado essencialmente qualitativo. Portanto, é um Dado Omisso Completamente ao Acaso (MCAR), sendo inviável a sua imputação por meio de cálculos.

Neste sentido, tendo que os dados omissos presentes nessa coluna representam uma baixa porcentagem de 4,44% do valor total do conjunto de dados, cuida-se de uma conjuntura que permite a a aplicação do método Listwise Deletion, no qual deletam-se os casos que contêm dados em falta, usando para a análise somente os casos que possui todos os dados presentes. Outro método que poderia ser utilizado nesta situação é o *Pairwise Deletion* contudo, como anteriormente mencionado, essa técnica dificulta a interpretação dos resultados obtidos, por simplesmente ignorar os dados omissos na análise.

VI. CONCLUSÕES

De tudo, tem-se que a qualidade do conhecimento extraído de determinado conjunto de dados depende, em muito, da qualidade e da integralidade do conjunto analisado. Destarte, a preparação dos dados para a sua análise é passo fundamental que não pode ser ignorado, vez que é neste momento que se identificam e tratam os dados omissos, que são uma grande fonte de adversidades na área.

Para solucionar esse contratempo utilizamos métodos para o tratamento dos dados em falta (*MDTs*), que devem ser empregadas observando às peculiaridades de cada caso, já que toda técnica possui vantagens e desvantagens. Outra recomendação é o teste dos métodos após a sua implementação para se assegurar que os dados omissos estão sendo tratados corretamente.

No caso concreto apresentado ao grupo, as soluções apresentadas no Notebook consideraram os padrões, os tipos e as percentagens dos dados omissos do conjunto. Como destacado, é recomendado também que se utilize ferramentas que permitam a visualização dos dados analisados, vez que são imprescindíveis na identificação padrões e tendências dos dados ausentes. Todas essas questões são importantes e devem ser consideradas no momento da seleção do melhor método para tratamento dos dados omissos.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Song, Q., & Shepperd, M. (2007). Missing Data Imputation Techniques. *International Journal of Business Intelligence and Data Mining*, 2(3), 261–291. https://doi.org/10.1504/IJBIDM.2007.015485.

[2] Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Prentice Hall.

[3] Mitra, S., & Acharya, T. (2019). Analysis and visualization of missing data in a survey. *Journal of Applied Statistics*, 46(10), 1793-1810. https://doi.org/10.1080/02664763.2018.1541353.

[4] Graham, J. W. (2009). *Missing data analysis*: Making it work in the real world. Guilford Press.

[5] Grus, J. (2015). *Data science from scratch*: First principles with Python (1st ed.). O'Reilly Media.

[6] Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

[7] Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data mining*: Practical machine learning tools and techniques (4th ed.). Morgan Kaufmann.

[8] Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

[9] Pereira, L. P., & Oliveira, C. (2023). Testes estatísticos para detecção de dados omissos. In S. Silva & J. Souza, *Métodos Estatísticos Aplicados* (pp. 67-68).

[10] Stef van Buuren (2018) *Flexible imputation of missing data*. Boca Raton: Chapman and Hall/CRC Press.

[11] Bilogur, A. (2018). Missingno: a missing data visualization suite. Journal of *Open Source Software*, 3(22), 547. https://doi.org/10.21105/joss.00547.

[12] Wilson, F. B. A. A. (2010). Tratamento de Valores Omissos Num Conjunto de Dados. http://monografias.uem.mz/handle/123456789/1809.

[13] Heymans, M.W., & Eekhout, I. (2019). *Applied Missing Data Analysis with SPSS and (R)Studio*. https://bookdown.org/mwheymans/bookmi/.

[14] Myrtveit, I., Stensrud, E. and Olsson, U.H. (2001) Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods, *IEEE Transactions on Software Engineering*, bol. 27, n. 11, p. 999–1013. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=965340.

[15] Scikit-Learn Developers (2022). *Imputation of missing values*. Consultado a 18/02/2023 em https://scikit-learn.org/stable/modules/impute.html#.

[16] Brown, M. L.; Kros, J. F. (2003). Data mining and the impact of missing data. *Industrial Management & Data Systems*, v. 103, n. 8, p. 611-621. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://eic.cefet-rj.br/ppcic/wp-content/uploads/2021/01/35-Leandro-Maia-Goncalves.pdf.

[17] Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple Imputation: A Flexible Tool for Handling Missing Data. JAMA, 314(18), 1966–1967. https://doi.org/10.1001/jama.2015.15281.

[18] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8(1), 1-37. https://doi.org/10.1186/s40537-021-00516-9.